Aman Chadha
amanc@stanford.edu

CS231n: Convolutional Neural Networks
for Visual Recognition

Project Proposal

Connecting Vision and Natural Language: End-to-End Transformer-Based Dense Video Captioning with Context Gating and Joint Ranking

Problem statement

Dense video captioning aims to localize and describe events in a video using natural language. This problem can be tackled by splitting it into two parts: (i) detecting event proposals and, (ii) generating captions for each of these events.

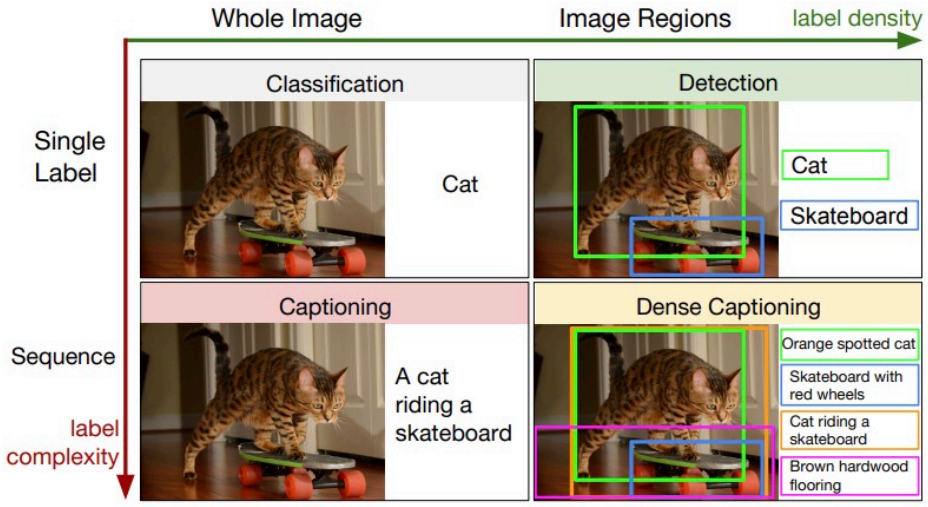

Figure 1. Dense Captioning (bottom right) originally introduced in [7] for images, involves generation of dense, rich annotations.

However, current methods suffer from several drawbacks:

- **Inability to capture long-range dependencies from both the past and the future for event proposal and captioning**: Most methods predominantly generate temporal event proposals in the forward direction and neglect future video context. For instance, while [3] uses a captioning module that utilizes the context from all the events from their proposal module, they only utilize past context for proposal prediction. To this end, [1] uses a multi-head attention mechanism while [2] and [4] utilize bidirectional LSTMs to effectively exploit both past and future contexts to make proposal predictions.
- **Using bidirectional-LSTMs rather than attention-based architectures:** In [5]**,** the authors show that attention-based architectures can outperform RNN-based techniques with significant performance improvements. [2] and [4] utilize bi-directional LSTMs. However, [1] uses a video encoder which is composed of multiple self-attention layers and uses multi-head attention for captioning.
- **End-to-end training v/s training with separation/alternation:** [3] trains the proposal and captioning model alternately. On the other hand, end-to-end training as in [1] enables the event proposal model to be directly influenced by the language description and thus generates accurate descriptions.

- **Limited captioning accuracy and diversity:** [1] uses a Masked Transformer for caption generation, while [2] uses Temporal Dynamic Attention. However, [4] deploys an ensemble of three different caption models, viz., Vanilla Caption Model [8], Temporal Attention Caption Model [9] and Topic Guided Caption Model [10], to improve captioning accuracy and diversity.
- **Lack of handling parallel multi-events:** Different events ending at (nearly) the same time are indistinguishable in the previous works, resulting in the same captions. [2] solves this problem by representing each event with an attentive fusion of hidden states from the proposal module and video contents (e.g., C3D features).

Proposed idea

This project presents a novel amalgamation of the best virtues of the above methods and thus addresses the above drawbacks. We propose an end-to-end transformer-based architecture that utilizes the following:

- Attention-based architectures for video encoding, event proposals and captioning. Specifically, multi-head attention for past/future context-aware proposal predictions and caption generation.
- End-to-end training to influence the event proposal model by the language description.
- Using an ensemble of caption models to generate accurate and diverse captions.
- Encoding the video input as a sequence of visual features to effectively handle parallel multi-events.
- Using a context gating mechanism to balance the contributions from the current event and its surrounding contexts dynamically.
- To improve result confidence, we plan to utilize the joint ranking technique in [2] at inference time to select high-confidence proposal-caption pairs by accounting for both the proposal score and caption confidence.

Dataset

ActivityNet Captions [3] and YouCookII [6] are the two of the largest datasets with temporal event segments annotated and described using natural language, which we would be training and evaluating our model on.

Results

For initial results, we plan to evaluate our captioning performance on the validation set of ActivityNet Captions using ground truth event proposals. We aim to evaluate the end-to-end system by analyzing captioning performance using learned event proposals on both the validation and test set of ActivityNet Captions. We plan to measure the precision of our captions using traditional evaluation metrics: BLEU, METEOR and CIDEr. If time permits, we plan to experiment with the recipe generation benchmark on the YouCookII validation set. We expect our results to demonstrate that the proposed model achieves improved results compared to the current methods owing to the above state-of-the-art techniques.

References

[1] Zhou, L., Zhou, Y., Corso, J. J., Socher, R., & Xiong, C. (2018). End-to-end dense video captioning with masked transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8739-8748).

[2] Wang, J., Jiang, W., Ma, L., Liu, W., & Xu, Y. (2018). Bidirectional attentive fusion with context gating for dense video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7190-7198).

[3] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017). Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision (pp. 706-715).

[4] Chen, S., Song, Y., Zhao, Y., Qiu, J., Jin, Q., & Hauptmann, A. (2018). RUC+CMU: System report for dense captioning events in videos. arXiv preprint arXiv:1806.08854.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[6] Zhou, L., Xu, C., & Corso, J. J. (2018, April). Towards automatic learning of procedures from web instructional videos. In Thirty-Second AAAI Conference on Artificial Intelligence.

[7] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4565-4574).

[8] Q. Jin, S. Chen, J. Chen, and A. Hauptmann. Knowing your- self: Improving video caption via in-depth recap. In Proceedings of the 2017 ACM on Multimedia Conference, pages 1906–1911. ACM, 2017.

[9] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In Proceedings of the IEEE international conference on computer vision, pages 4507–4515, 2015.

[10] S. Chen, J. Chen, and Q. Jin. Generating video descriptions with topic guidance. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pages 5–13. ACM, 2017.